Generalized MDS for Data Exploration, Discriminant Analysis, Clustering and Visualization

David A. Johannsen*
Dahlgren Division of the Naval Surface Warfare Center
Advanced Computation Technology Group Code B10

Jeff Solka†
Dahlgren Division of the Naval Surface Warfare Center
Advanced Computation Technology Group Code B10

April 29, 2005

Abstract

Multidimensional scaling (MDS) is a well established and frequently used tool for data dimensionality reduction and visualization. More recent work in dimensionality reduction has begun exploration of such areas as "manifold discovery." Our work extends the methods of MDS to a more general setting in order to more fully exploit discovered manifold structure in data.

1 Introduction

Multidimensional scaling (MDS) is a well established and frequently used tool for data dimensionality reduction and visualization (see, e.g., [1] and [2] for two standard references). This methodology determines a configuration of points in a low-dimensional space which best approximates the interpoint distances of the original data (in some space of higher dimensionality). The solution to the MDS problem is often determined as that configuration of points in a low–dimensional target space which minimizes a function of the differences of the interpoint distances of the original data and the configuration (the function is usually some sort of sum of squared differences between the interpoint distances).

Some of the more recent work in dimensionality reduction presumes the existence of an embedded (sub)manifold on which the data is constrained to lie. This assumption is based on the belief that there are non-linear and unknown relations between the observed/measured features (i.e., that, in practice, the data can be adequately described by a reasonably small number of independent parameters). The current tools of Isometric Mapping (IsoMap) [6] and Locally Linear Embedding (LLE) [4] attempt to preserve this manifold structure, locally. However, the final step in both of these algorithms is the selection of a configuration of points in Euclidean space of some low dimension, $\mathbb{R}^k$. It is in this low-dimensional Euclidean space where the practitioner performs visualization and metric exploitation of the data (i.e., classification, clustering, etc.).

Our work represents first steps to expanding the variety of spaces to which the final projection can be done. We call our work "generalized MDS," and in it we have expanded the class of spaces that one can use as the "target space" for MDS. Specifically, beyond $\mathbb{R}^2$, we can perform MDS to $\mathbf{S}^2$ and $\mathbb{H}^2$, the spaces of constant sectional curvature $+1$ and -1, respectively. We also can perform generalized MDS

*Tel. 540-653-2737 email david.johannsen@navy.mil
†Tel. 540-653-1982 email jeffrey.solka@navy.mil

to closed and orientable surfaces, endowed with their (unique) compatible constant curvature metric, i.e., 2-dimensional Euclidean and Hyperbolic space forms.

2 A Quick Review of Some Geometry

In two dimensions, there is a convenient classification (up to homeomorphism) of compact orientable surfaces. Specifically, the genus (the number of "handles" attached to a sphere) provides the topological invariant (see [3] for a presentation of the classification of closed surfaces). Furthermore, via the classical Gauss-Bonnet Theorem for surfaces, the genus also determines a unique constant curvature metric with which this surface can be endowed:

$$\int_M K \; dS = 2\pi(2-2g)$$

(more correctly, the Gauss-Bonnet theorem proves the uniqueness of the metric and the Poincaré Polyhedron Theorem provides the existence).

When endowed with their compatible constant curvature metric, there is a natural interpretation of these surfaces as homogeneous space forms; i.e., quotients of $\mathbb{R}^2$, $\mathbf{S}^2$ or $\mathbb{H}^2$ by discrete subgroups of their isometry groups (for a more detailed treatment of space forms, see [7]).

The point of the above is the following: Suppose that one has determined that a data set is intrinsically 2-dimensional. Suppose further that one has "discovered" manifold structure in the data (e.g., by fitting a simplicial complex to the data). Then there is a natural choice of space form to which one should perform MDS in order to obtain dimensionality reduction and data regularization — the constant curvature space which is homeomorphic to the discovered manifold structure. Even if effective "manifold discovery" remains an elusive goal, the generalized MDS procedures give new tools in the kit of practitioners of data exploration and visualization.

To this point, our work has concentrated on MDS to 2-dimensional Riemannian manifolds (i.e., closed and orientable surfaces endowed with constant curvature metrics). We mention that our work can be extended to higher dimensions in a relatively straightforward manner. First, Thurston's Geometrization Theorem (perhaps being a bit premature in calling this a theorem) gives a classification of closed and orientable 3-manifolds. Thus, in three dimensions we have an exhaustive enumeration of manifolds similar to that used in our 2-dimensional work (albeit a bit more complicated, involving decomposition of the 3-manifold into pieces, each of which supports one of eight model geometries). Now it is a well known (and elementary) result that there can be no "classification" of manifolds in dimensions four and higher (essentially because the "word problem" is unsolvable, e.g., [5]). However, the methodologies that we have produced admit obvious extension to four and higher dimensional space forms; it is just now the case that such space forms are far from including the totality of all possible closed and orientable manifolds. However, we will say that such space forms (i.e., quotients of the three constant curvature spaces by subgroups of their isometry groups) still represents a very rich class of spaces.

3 The Generalized MDS Procedure

Non-classical MDS amounts to minimization of a function whose value is determined by the relative positions of a configuration of points. From the original data, one can compute an interpoint distance matrix $\hat{D}$. This interpoint distance matrix is most usually computed using the usual Euclidean metric of the ambient Euclidean space, $\mathbb{R}^n$. However, it has been shown that the graph distance of IsoMap is an arbitrarily good approximation to the induced (submanifold) metric of the (sub)manifold on which the data are presumed to lie (i.e., as the density of the sampling increases the neighborhood graph metric converges asymptotically to the submanifold metric). Thus, since we presume that the data has been sampled from an embedded sub-manifold, we have used the IsoMap algorithm to compute our ambient interpoint distance matrix. We remark that our original intention was to fit a simplicial complex to the data

(i.e., to perform "manifold discovery"), and use the natural PL metric of the complex to compute the initial interpoint distance matrix. However we have neither been able to locate an existent algorithm which performs acceptably nor have we been able to devise one of our own.

Once one has selected a target space for the MDS, then the minimization of the functional becomes a standard problem of minimizing a function on a Riemannian manifold. We perform this minimization via steepest descent. That is, even in this more general setting, it is still true that minus the gradient is a descent direction for a function, $f : M \to \mathbb{R}$, where M is a Riemannian manifold. One must simply determine the explicit form of the exponential map $\exp : T_{x_k}M \to M$ at the current iterate, x_k, in order to perform the backtracking.

The chief difficulty in the case of finding a configuration of points in a hyperbolic space form is explicit determination of the exponential map, $\exp : T_pM \to M$, where M is a hyperbolic space form and T_pM is the tangent space to the manifold at the point p. Recall that there is no isometric embedding $\mathbb{H}^2 \hookrightarrow \mathbb{R}^3$ (i.e., no realization of the hyperbolic plane in $\mathbb{R}^3$ so that the induced metric is the metric of constant curvature -1). Thus, given a descent direction v for a function f at a point p one needs to find the geodesic through p with tangent vector v (this is the exponential map and is a diffeomorphism between a ball centered at the origin of the tangent space and some neighborhood of p). In order to perform the minimization, we have determined the explicit form of this mapping in the Poincaré model of hyperbolic space.

The second hurdle comes when one considers nontrivial hyperbolic space forms. As mentioned above, this is equivalent to performing MDS to a surface of genus $g > 1$ endowed with a constant curvature metric. Such a surface can be described as the space obtained by identifying edges of a regular hyperbolic $4g$-gon, much as the flat torus is obtained by identifying the edges of regular (Euclidean) rectangle. More generally, one can think of this hyperbolic $4g$-gon as a quotient of hyperbolic space by the action of a subgroup of the full isometry group, i.e., the subgroup that is generated by $2g$ isometries (i.e., the face pairings). If we denote these isometries by $\{\phi_1, \ldots, \phi_{2g}\}$ and the free abelian group that these isometries generate by $G =< \phi_1, \ldots, \phi_{2g} >$, then the distance between two points, x and y on the genus g surface is given by

$$\min_{\phi \in G} d(x, \phi(y)),$$

where $d(\cdot, \cdot)$ is the distance function induced by the constant curvature metric on $\mathbb{H}^2$. We have been able to explicitly determine these isometries and thus can compute distances using the appropriate metric.

4 Exploitation and Results

We now describe the final phase of our work. Once one has applied the generalized MDS and determined a configuration of points in a surface which is endowed with its constant curvature Riemannian metric (as in the preceding section), one can explore the structure of the data. Moreover, the manifold with its intrinsic metric seems to us to be the natural setting in which to do nearest neighbor classification and (metric) clustering of the data. That is, if the assumption is valid that the data resides on some low-dimensional manifold, then the intrinsic metric seems to be the natural setting for distance computations (and not the metric of an ambient Euclidean space). Thus, one of our basic questions of interest is whether classifier and clustering performance improves when one uses the appropriate target space and its intrinsic metric (for interpoint distance computations).

If the data is (intrinsically) 2-dimensional, then the algorithms that we have described have obvious benefit for visualization. Since the closed and oriented 2-dimensional manifolds can be embedded in (Euclidean) $\mathbb{R}^3$, one can use any number of available plotting programs to display the surface and the projected data. We believe that this will often give a much truer picture of the structure inherent in the data than a projection to a region of the Euclidean plane.

We are currently applying these MDS techniques to computer user profiling data. We soon will have results that can be presented. At a minimum, the explicit determination of generalized MDS mentioned in the preceding section is of theoretical interest.

5 Future Directions

There is an obvious class of 2-dimensional spaces to which to perform MDS, namely the three homogeneous and isotropic spaces, $\mathbb{R}^2$, $\mathbf{S}^2$, and $\mathbb{H}^2$, and their quotients (as these quotients exhaust "all" closed and orientable surfaces). It is unfortunate that one cannot apply this same methodology in arbitrary dimension. However, there is still hope in three dimensions. It is known that there are only eight 3-dimensional (locally homogeneous) model geometries. Moreover, Perelman's proof of the Geometrization Conjecture gives a classification of closed and orientable 3-dimensional manifolds. Thus, we would like to extend our work to the case of 3-dimensional manifolds. Above dimension three, there is no hope of any sort of complete classification of closed and orientable manifolds. Moreover, even in dimension two, there is not yet a classification of all possible hyperbolic space forms. However, one may select some reasonably nice collection of Riemannian manifolds to which to perform MDS and subsequent exploitation (for example, some distinguished class of quotients of the three homogeneous and isotropic spaces).

References

[1] I. Borg, and P. Groenen, *Modern Multidimensional Scaling, Theory and Applications*, Springer, New York, 1997.

[2] T. F. Cox, and M. A. A. Cox, *Multidimensional Scaling*, 2nd ed., Chapman and Hall/CRC, Boca Raton, 2001.

[3] W. S. Massey, *A Basic Course in Algebraic Topology*,Grad. Texts in Math. vol. 127, Springer-Verlag, New York, 1991.

[4] S. T. Roweis, and L. K. Saul, *Nonlinear dimensionality reduction by locally linear embedding*, Science 290, pp 2323 – 2326, 22 December 2000.

[5] J. Stillwell, *Classical Topology and Combinatorial Group Theory*, Grad. Texts in Math. vol. 72, Springer-Verlag, New York, 1993.

[6] J. B. Tenenbaum, V. de Silva, and J. C. Langford, *A global geometric framework for nonlinear dimensionality reduction*, Science 290, pp. 2319 – 2323, 22 December 2000.

[7] J. A. Wolf, *Spaces of Constant Curvature*, 5th ed., Publish or Perish, Inc., Wilmington, 1984.